PCT WORLD INTELLECTUAL PROPERTY ORGANIZATION International Bureau

B1

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : C12Q 1/68 — A2

(11) International Publication Number: WO 99/07896

(43) International Publication Date: 18 February 1999 (18.02.99)

(21) International Application Number: PCT/US98/16548

(22) International Filing Date: 7 August 1998 (07.08.98)

(30) Priority Data:
60/054,887 7 August 1997 (07.08.97) US

(63) Related by Continuation (CON) or Continuation-in-Part (CIP) to Earlier Application
US
Filed on 60/054,887 (CIP) 7 August 1997 (07.08.97)

(71) Applicant *(for all designated States except US)*: CURAGEN CORPORATION [US/US]; 11th floor, 555 Long Wharf Drive, New Haven, CT 06511 (US).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: ROTHBERG, Jonathan, M. [US/US]; 1701 Moose Hill Road, Gilford, CT 06437 (US). DEEM, Michael, W. [US/US]; 11136 Ophir Drive, Los Angeles, CA 90024 (US). SIMPSON, John, W. [US/US]; 23 Woodland Road B4, Madison, CT 06443 (US).

(74) Agent: ELRIFI, Ivor, R.; Mintz, Levin, Cohn, Ferris, Glovsky and Popeo, P. C., One Financial Center, Boston, MA 02111 (US).

(81) Designated States: AL, AM, AT, AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, DE, DK, EE, ES, FI, GB, GE, GH, GM, HR, HU, ID, IL, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR, TT, UA, UG, US, UZ, VN, YU, ZW, ARIPO patent (GH, GM, KE, LS, MW, SD, SZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).

Published
Without international search report and to be republished upon receipt of that report.

(54) Title: DETECTION AND CONFIRMATION OF NUCLEIC ACID SEQUENCES BY USE OF OLIGONUCLEOTIDES COMPRISING A SUBSEQUENCE HYBRIDIZING EXACTLY TO A KNOWN TERMINAL SEQUENCE AND A SUBSEQUENCE HYBRIDIZING TO AN UNDENTIFIED SEQUENCE

(57) Abstract

The present invention discloses a methodology which is directed to providing positive confirmation that nucleic acids, possessing putatively identified sequences predicted to generate observed GeneCalling™ signals, are actually present within the sample from which the signal was originally derived. The putatively identified nucleic acid fragment within the sample possesses 3'– and 5'–ends with known terminal subsequences, said method comprising: contacting said nucleic acid fragments in said sample in amplifying conditions with (i) a nucleic acid polymerase; (ii) "regular" primer oligonucleotides having sequences comprising hybridizable portions of said known terminal subsequences; and (iii) a "poisoning" oligonucleotide primer, said poisoning primer having a sequence comprising a first subsequence that is a portion of the sequence of one of said known terminal subsequences and a second subsequence that is a hybridizable portion of said putatively unidentified sequence which is adjacent to said one known terminal subsequence, wherein nucleic acids amplified with said poisoning primer are distinguishable upon detection from nucleic acids amplified with said nucleic acids amplified only with said regular primers; separating the products of the contacting step; and the detecting sequence is confirmed if the nucleic acids amplified with said poisoning primer are detected.

DETECTION AND CONFIRMATION OF NUCLEIC ACID SEQUENCES BY USE OF OLIGONUCLEOTIDES COMPRISING A SUBSEQUENCE HYBRIDIZING EXACTLY TO A KNOWN TERMINAL SEQUENCE AND A SUBSEQUENCE HYBRIDIZING TO AN UNDENTIFIED SEQUENCE

RELATED APPLICATIONS AND GRANT SUPPORT

This application claims priority to United States Provisional Patent Application Serial No. 60/054,887 originally filed on August 7, 1997, which is entitled "METHOD AND APPARATUS FOR IDENTIFYING, QUANTIFYING, AND CONFIRMING DNA SEQUENCES IN A SAMPLE WITHOUT SEQUENCING", and is hereby incorporated in its entirety by reference herein.

The invention disclosed herein was made utilizing United States Government support under Grant Number 70NANB5H1036 awarded by the National Institute of Standards and Technology. Accordingly, the United States Government has certain rights in the invention.

FIELD OF THE INVENTION

The field of the invention is DNA sequence classification, identification or determination, and quantification; more particularly it is the quantitative classification, comparison of expression, or identification of preferably all DNA sequences or genes in a sample without performing any associated sequencing.

BACKGROUND OF THE INVENTION

As molecular biological and genetics research have advanced, it has become increasingly clear that the temporal and spatial expression of genes plays a vital role in processes occurring in both health and in disease. Moreover, the field of biology has progressed from an understanding of how single genetic defects cause the traditionally recognized hereditary disorders (e.g., the thalassemias), to a realization of the importance of the interaction of multiple genetic defects in concert with various environmental factors in the etiology of the majority of the more complex disorders, such as neoplasia.

For example, in the case of neoplasia, recent experimental evidence has demonstrated the key causative roles of multiple defects in several pivotal genes causing their altered expression. Other complex diseases have been shown to have a similar etiology. Therefore, the more complete and reliable a correlation which can be established between gene expression and

The oligonucleotide probes utilized in colony selection protocols for unknown gene(s) are synthesized to hybridize, preferably, only with the cDNA for the gene of interest. One method of achieving this specificity is to start with the protein product of the gene of interest. If a partial sequence (i.e., from a peptide fragment containing 5 to 10 amino acid residues) from an active region of the protein of interest can be determined, a corresponding 15 to 30 nucleotide (nt.) degenerate oligonucleotide can be synthesized which would code for this peptide fragment. Thus, a collection of degenerate oligonucleotides will typically be sufficient to uniquely identify the corresponding gene. Similarly, any information leading to 15-30 nt. subsequences can be used to create a single gene probe.

Another existing method, which searches for a known gene in cDNA or gDNA prepared from a tissue sample, also uses single-gene or single-sequence oligonucleotide probes which are complementary to unique subsequences of the already known gene sequences. For example, the expression of a particular oncogene in sample can be determined by probing tissue-derived cDNA with a probe which is derived from a subsequence of the oncogene's expressed sequence tag. The presence of a rare or difficult to culture pathogen (e.g., the TB bacillus) can also be determined by probing gDNA with a hybridization probe specific to a gene possessed by the pathogen. Similarly, the heterozygous presence of a mutant allele in a phenotypically normal individual, or its homozygous presence in a fetus, may be determined by the utilization of an allele-specific probe which is complementary only to the mutant allele. See e.g., Guo, N.C., *et al.* 1994. *Nucleic Acid Research* 22:5456-5465).

current state of art

Currently, all of the existing methodologies which utilize single-gene probes, if applied to determine all of the genes expressed within a given tissue sample, would require many thousands to tens-of-thousands of individual probes. It has been estimated that a single human cell typically expresses approximately 5,000 to 15,000 genes *simultaneously,* and that the most complex types of tissues (e.g., brain tissue) can express up to one-half of the total genes contained within the human genome. See Liang, *et al.* 1992. Differential Display of Eukaryotic Messenger RNA by Means of the Polymerase Chain Reaction. *Science* 257:967-971. It is obvious that an screening methodology which requires such a large number of probes is clearly far too cumbersome to be economic or, even practical.

In contrast, another class of existing methods, known as sequencing-by-hybridization ("SBH"), utilize combinatorial probes which are not gene specific. See e.g., Drmanac, *et al.* 1993. *Science* 260:1649-1652; U.S. Patent No. 5,202,231 to Drmanac, *et al.* An exemplar implementation of SBH for the determination of an unknown gene requires that a single cDNA

("PCR") to amplify DNA subsequences of various lengths, which are then defined by their being between the annealing sites of arbitrarily selected primers. Polymerase chain reaction method and apparatus are well known. See, e.g., United Sates patents 4,683,202; 4,683,195; 4,965,188; 5,333,675; each herein fully incorporated by reference. Ideally, the pattern of the lengths observed is characteristic of the specific tissue from which the library was originally prepared. Typically, one of the primers utilized in differential display is oligo(dT) and the other is one or more arbitrary oligonucleotides which are designed to hybridize within a few hundred base pairs (bp.) of the homopolymeric poly-dA tail of a cDNA within the library. Thereby, upon electrophoretic separation, the amplified fragments of lengths up to a few hundred base pairs should generate bands which are characteristic and distinctive of the sample. In addition, changes in gene expression within the tissue may be observed as changes in one or more of the cDNA bands.

In the differential expression methodology, although characteristic electrophoretic banding patterns develop, no attempt is made to quantitatively "link" these patterns to the expression of particular genes. Similarly, the second arbitrary primer also cannot be traced to a particular gene due to the following reasons. First, the PCR process is less than ideally specific. One to several base pair mismatches are permitted by the lower stringency annealing step which is typically utilized in this methodology and are generally tolerated well enough so that a new chain can actually be initiated by the *Tag* polymerase often used in PCR reactions. Secondly, the location of a single subsequence (or its absence) is insufficient to distinguish all expressed genes. Third, the resultant bp.-length information (i.e., from the arbitrary primer to the poly-dA tail) is generally not found to be characteristic of a sequence due to: (*i*) variations in the processing of the 3'-untranslated regions of genes, (*ii*) variation in the poly-adenylation process and (*iii*) variability in priming to the repetitive sequence at a precise point. Therefore, even the bands which are produced often are smeared by numerous, non-specific background sequences.

Moreover, known PCR biases towards nucleic acid sequences containing high G+C content and short sequences, further limit the specificity of this methodology. In accord, this technique is generally limited to the "fingerprinting" of samples for a similarity or dissimilarity determination and is precluded from use in quantitative determination of the differential expression of identifiable genes.

(motivation)

Thus, in conclusion, the existing methodologies utilized for gene or DNA sequence classification and determination are in need of improvement with respect to their ability to

conditions with (*i*) a nucleic acid polymerase; (*ii*) "regular" primer oligonucleotides having sequences comprising hybridizable portions of said known terminal subsequences; and (*iii*) a "poisoning" oligonucleotide primer, said "poisoning" primer having a sequence comprising a first subsequence that is a portion of the sequence of one of said known terminal subsequences and a second subsequence that is a hybridizable portion of said putatively unidentified sequence which is adjacent to said one known terminal subsequence, wherein nucleic acids amplified with said "poisoning" primer are distinguishable upon detection from nucleic acids amplified with said nucleic acids amplified only with said regular primers; separating the products of the contacting step; and the detecting sequence is confirmed if the nucleic acids amplified with said "poisoning" primer are detected.

The present invention further provides that: (*i*) the regular PCR primers are labeled and, preferably, said "poisoning" primer is unlabeled; (*ii*) the regular PCR primers are labeled and the "poisoning" primer is labeled in a detectably different manner so as to allow its differentiation from any other label utilized in the amplification reaction; or (*iii*) the regular PCR primers are unlabeled and the poisoning primer is labeled and, optionally, wherein the step of detecting said separated products further comprises confirming said putatively identified sequence if said nucleic acid fragment with a putatively identified sequence is not detected. In the preferred embodiment of the present invention the regular PCR primers are labeled and, preferably, said "poisoning" primer is unlabeled.

It is an object of this invention to provide a methodology for the rapid, economical, quantitative, and highly specific determination or classification of DNA sequences, in particular genomic DNA (gDNA) or complementary DNA (cDNA) sequences, in either arrays of single sequence clones or mixtures of sequences such as can be derived from tissue samples, without actually sequencing the DNA. Thereby, the aforementioned deficiencies within the background arts are greatly mitigated. This objective is realized by generating a plurality of distinctive and detectable signals from the DNA sequences in the sample being analyzed. Preferably, all the resultant signals taken together have sufficient discrimination and resolution so that each particular DNA sequence contained within a sample may be individually classified by the particular signals it generates, and with reference to a database of all DNA sequences possible in the sample, individually determined. The intensity of the signals indicative of a particular DNA sequence depends, preferably, on the amount of that DNA present. Alternatively, the signals together can classify a predominant fraction of the DNA sequences into a plurality of sets of approximately no more than two to four individual sequences.

accomplished by knowing the sequence of each clone and/or by determining the length (either measured or physical) of the recognized sequences.

The generated signals are then analyzed together with DNA sequence information stored within sequence databases utilizing computer implemented experimental analysis methods to: (*i*) identify individual genes and (*ii*) establish their quantitative presence within the sample. The target subsequences are chosen by further computer implemented experimental design methods of the present invention such that their presence or absence, as well as their relative distances when present, yield a maximum amount of information for classifying or determining the DNA sequences to be analyzed.

By use of this methodology, it is possible to have orders of magnitude fewer probes than there are DNA sequences to be analyzed, and it is further possible to have considerably fewer probes than would be present in combinatorial libraries of the same length as the probes used in this invention. The target subsequences have a preferred probability of occurrence in a sequence (typically between 5% and 50%). In the preferred embodiment, it is preferred that the presence of one probe in a DNA sequence to be analyzed is independent of the presence of any other probe. Preferably, target subsequences are chosen based on information in relevant DNA sequence databases that characterize the sample. A minimum number of target subsequences may be chosen to determine the expression of all genes in a tissue sample (hereinafter "tissue mode"). Alternatively, a smaller number of target subsequences may be chosen to quantitatively classify or determine only one or a few sequences of genes of interest, for example oncogenes, tumor suppressor genes, growth factors, cell, cycle genes, cytoskeletal genes, and the like (hereinafter "query mode").

The preferred embodiment of this detection methodology, quantitative expression analysis (hereinafter referred to as "GeneCalling™") generates signals which comprise both the target subsequence presence and a representation of the length in base pairs between adjacent target subsequences via the measurement of the results of recognition reactions on cDNA (or gDNA) mixtures. A detailed disclosure of the GeneCalling™ methodology may be found in PCT/US96/17159, published as WO97/15690, herein incorporated by reference, which is entitled "METHOD AND APPARATUS FOR IDENTIFYING, QUANTIFYING, AND CONFIRMING DNA SEQUENCES IN A SAMPLE WITHOUT SEQUENCING." Most importantly, this methodology *does not* require the insertion of the cDNA into a vector so as to create individual clones in a library. It is well known within the relevant fields that the creation of these cDNA libraries is time consuming, costly, and introduces bias into the process, as it requires the cDNA

useful to be able to assay the genetic makeup and expression of a tissue sample. More specifically, the presence and expression of certain genes or their particular alleles are prognostic or risk factors for disease (including disorders). Several examples of such diseases are found among the neurodegenerative diseases, such as Huntington's disease and ataxia-telangiectasia. Several cancers (e.g., neuroblastoma) can now be quantitatively linked to specific genetic defects. Finally, gene expression can also determine the presence and classification of those foreign pathogens which are difficult or impossible to culture *in vitro* but which nevertheless express their own unique genes.

Similarly, disease progression is reflected in changes in genetic expression of an affected tissue. For example, expression of particular tumor promoter genes and lack of expression of particular tumor suppressor genes is now known to correlate with the progression of certain tumors from normal tissue, to hyperplasia, to cancer *in situ*, and finally, to metastatic cancer. The return of a cell population to a "normal" pattern of gene expression (e.g., through the use of anti-sense oligonucleotide technology), can correlate with tumor regression. The quantificati n of gene expression in a cancerous tissue can assist in staging and classifying this disease, as well as providing a basis to chose and guide therapy. Accurate disease classification and taging or grading using gene expression information can assist in choosing initial therapies that are increasingly more precisely tailored to the precise disease process occurring in the particular patient. Gene expression information can then track disease progression or regression, and such information can assist in monitoring the success or changing the course of an initial therapy. A favored therapy is one which results in a regression of an abnormal pattern of gene expression in an individual towards "normality," while a therapy which has little effect on gene expression (i.e., its abnormal progression) may be modified or discontinued. Such monitoring of gene expression is now useful for cancers and will become useful for an increasing number of other diseases, such as diabetes and obesity.

In order to facilitate the utilization of the present invention for the quantitative detection, confirmation and monitoring of such differential gene expression in patients with the aforementioned diseases, it is envisioned that the GeneCalling™/oligo-poisoning methodologies will be incorporated into a unitized "kit" form. This will enable the researcher or health care provider to rapidly and accurately assess such differential gene expression in the most efficacious manner possible. For example, the kit may utilize non-radioactive labeling of the PCR amplification probes and "pre-cast" electrophoresis gels to ameliorate some of the difficulties

DESCRIPTION OF THE FIGURES

The present invention may be better understood and its advantages appreciated by those individuals skilled in the relevant arts by referring to the accompanying drawings wherein:

Figure 1 is an illustration of the DNA primers utilized for a PCR amplification-mediated embodiment of the GeneCalling™ methodology.

Figure 2 is a flow chart illustrating the various steps utilized in the oligo-poisoning methodology of the present invention as applied to the conformation of the results obtained from GeneCalling.™

Figure 3, Panels A & B are schematic diagrams illustrating the preferred construction of the "poisoning" primers utilized in the oligo-poisoning methodology of the present invention as applied to the conformation of the results obtained from GeneCalling.™

Figure 4 provides the nucleotide sequence of a 2493 bp. cDNA generated from the mRNA encoded by the Human Complement Component 1, Subcomponent r (C1r) gene. The bold sequence illustrates the 319 bp. subsequence generated by digestion of the C1r cDNA with the REs BspH1 and EcoR1. This cDNA was utilized with the oligo-poisoning conformation methodology of the present invention.

Figure 5, Panels A & B are electropherograms of the up and down traces generated by GeneCalling™ PCR amplification reactions with the C1r cDNA. These reactions utilized either the "J" primer (complementary to the BspH1 recognition sequence) or "R" primer (complementary to the EcoR1 recognition sequence).

Panel A: Electropherogram of GeneCalling™ up trace.

Panel B: Electropherogram of the GeneCalling™ down trace.

Figure 6, Panels A & B are electropherograms of the up and down traces generated by PCR amplification of the GeneCalling™ reactions with C1r cDNA utilizing "poisoning" oligomer primers.

Panel A: Electropherogram of the oligo-poisoned GeneCalling™ up trace.

various steps in the preferred embodiment of GeneCalling™ methodology (i.e., the restriction endonuclease digestion/ligation/ amplification-based protocol) may be summarized as follows:

Step 1: complementary DNA (cDNA) synthesis

Step 2: The resulting cDNA fragments are digested utilizing two different restriction endonucleases (RE) which, preferably, recognize only rare, 6-8 bp. RE-recognition sequences.

Step 3: Ligation of oligonucleotide "adapters" to the digested cDNA fragments. Two different adapters are utilized, with each adapter being complementary to the sequences of one of the two RE recognition sites.

Step 4: PCR amplification is performed utilizing labeled primers which are complementary to the two adapters ligated to the digested cDNA fragments.

Step 5: The reaction products of the PCR amplification are then electrophoresed to observe the electrophoretic mobility patterns of the individual fragments. These mobility patterns are then utilized to construct an electropherogram.

Step 6: From the electrophoretic mobility and electropherogram the sizes of the individual fragments of interest are identified, and a computer DNA sequence database is then searched to generate a list of putative gene "identities" for these aforementioned fragments.

Thus, the GeneCalling™ methodology is performed by hybridizing the sample with one or more labeled probes, wherein each probe recognizes a different "target" nucleotide subsequence or a different set of "target" nucleotide subsequences. The target subsequences utilized in the GeneCalling™ methodology are, preferably, optimally chosen by the computer implemented methods of this invention in view of DNA sequence databases containing sequences likely to occur in the sample to be analyzed. In respect to the analysis of human

While the GeneCalling™ methodology is preferred for classifying and determining sequences contained within a sample comprised of a mixture of cDNAs, but it is also adaptable those samples which contain a single cDNA moiety. Typically, enough pairs of target subsequences can be chosen so that sufficient distinguishable signals may be generated so as to allow the determination of one, to all of the sequences contained within the sample mixture. For example, in a first possible scenario, any pair of target subsequences may occur more than once in a single DNA molecule to be analyzed, thereby generating several signals with differing lengths from one DNA molecule. In a second scenario, even if a pair of target subsequences occurred only once within two different DNA molecules to be analyzed, the lengths between the probe hybridizations may differ, and thus distinguishable hybridization signals may be generated.

In the preferred PCR-mediated GeneCalling™ methodology, a suitable collection of target subsequences is chosen via computer-implemented methods and PCR primers, preferably labeled with fluorescent moieties, are synthesized to hybridize with these aforementioned target subsequences. Advances in fluorescent labeling techniques, in optics, and in optical sensing currently permit multiply-labeled DNA fragments to be differentiated, even if they spatially-overlap (i.e., occupy the same "spot" on a hybridization membrane or a band within a gel). See Ju, T., *et al.*, 1995. *Proc. Natl. Acad Sci. USA* 92:4347-4351. Accordingly, the results of several GeneCalling™ reactions may be multiplexed within the same gel lane or filter spot. The primers are designed to reliably recognize short subsequences while achieving a high specificity in the PCR amplification step. Utilizing these primers, a minimum number of PCR amplification steps amplifies those fragments between the primed subsequences existing in DNA sequences in the sample, thereby recognizing the target subsequences. The labeled, amplified fragments are then separated by gel electrophoresis and detected.

GeneCalling™ may be performed in either a "query mode" or in a "tissue mode." In query mode, the focus is upon the determination of the expression of a limited number of genes of interest and of known sequence (e.g., those genes which encode oncogenes, cytokines, and the like). A minimal number of target subsequences are chosen to generate signals, with the goal that each of the limited number of genes is discriminated from all the other genes likely to occur in the sample by at least one unique signal. Conversely, in tissue mode, the focus is upon the determination of the expression of as many as possible of the genes expressed in a tissue or other sample, without the need for any prior knowledge of their expression. In the tissue mode, target subsequences are optimally chosen to discriminate the maximum number of sample DNA sequences into classes comprising one, or preferably at-most a few sequences. Ideally, sufficient

and 502. It should be noted that Although Component 503 is optional, it can be utilized to improve the specificity of the first low stringency annealing step during PCR amplification, and thereby mitigate the production of false positive amplification products. Component 502 is a sequence which is complementary to the subsequence which primer 501 is designed to recognize. Component 502 is typically 4-8 bp in length. Component 504 is a 10-20 bp sequence chosen so the final primer does not hybridize with any native sequence in the cDNA sample to be analyzed; that is, primer 501 does not anneal with any sequence known to be present in the sample to be analyzed. The sequence of component 504 is also chosen so that the final primer has a melting temperature (T_m) above 50°C and preferably, above 68°C.

Use of primer 501 in the PCR amplification involves a first annealing step, which allows the 3' terminal Component 502 to anneal to its target subsequence in the presence of Component 504, which may not hybridize. Preferably, this annealing step is performed at a temperature (generally between 36°C and 44°C) which is empirically determined so as to maximize the reproducibility of the resulting hybridization signal pattern. The DNA concentration is approximately 10 ng/50 ml and is similarly determined to maximize reproducibility. Once annealed, the 3'-terminus serves as the primer elongation point for the subsequent first elongation step. The first elongation step is preferably performed at 72°C for 1 minute. If stringency conditions are such that exact complementarity is not required for primer annealing, false positive signals(i.e., signals resulting from inexact recognition of the target subsequence) may be generated. Subsequent cycles utilize high temperature, high stringency annealing steps. The high stringency annealing steps ensure annealing of the entire primer to the fragment sequences of interest. Preferably, these PCR cycles alternate between a 65°C annealing step and 95°C melting step, with each step lasting for 1 minute.

As previously discussed, optional Component 503 may be utilized to improve the specificity of the first low stringency annealing step, and thereby minimize the production of false positive bands. Component 503 may possess the generic formula -$(N)^j$-; where N is any nucleotide and j is typically between 2 and 4, preferably 2. Use of all possible Component 503 configurations results in a degenerate set of primers which have a 3'-terminus subsequence which is effectively "j" nucleotides longer in length than that of the target subsequence. These extended complementary end sequences have improved hybridization specificity. Alternately, Component 503 can be where N is a "universal" nucleotide and j is typically between 2 and 4, preferably 3 or 4. A universal nucleotide (e.g., inosine), is one which capable of forming base pairs with any other naturally occurring nucleotide. In this alternative, single primer 501 has a

Step 2: Phenol/CHCl, extract the mixture 2 times, and purify it on a Centricon 100 (Millipore Corporation; Bedford, MA) using water as the filtrate.

Step 3: Bring the end volume of the cDNA to 50 4 (starting with 10 ng RNA/µl).

Step 4: Set up the following PCR Reaction:

Component	Volume
cDNA (-10 ng/µl)	1 µl
10X PCR Buffer	2.5 µl
25 mM $MgCl_2$	1.5 µl
10 mM dNTPs	0.5 µl
20 pM/µl primer 1	2.5 µl
20 pM/µl primer 2	2.5 µl
Taq Poly. (5 U/µl)	0.2 µl
Water	14.3 µl

Step 5: One low stringency cycle with the profile:
40°C for 3 minutes (annealing)
72W for 1 minute (extension)

Step 6: Cycle using the following profile:
95°C for 1 minute (15-30 times):
95°C for 30 seconds
50°C for 1 minute
72°C for 1 minute
72°C for 5 minutes

Step 7: 4°C hold.

Step 8: Samples are precipitated, resuspended in denaturing loading buffer,

compared with the putatively identified sequence. Hence, confirmation is obtained if a fragment exists within the sample which possess a central subsequence having a sequence which is (at a minimum) homologous to a portion of the putatively identified sequence.

Nucleic acids possessing this generic structural motif are, preferably, produced according to the GeneCalling™ methods of this invention. As will be described in Section 2(a) *supra*, a preferred embodiment of the oligo-poisoning methodology is utilized in confirming that a specific sequence, obtained through the use of a nucleic acid sequence computer database, which has been predicted to generate a particular GeneCalling™ signal is, in actuality, generating the signal. Nonetheless, this embodiment of the oligo-poisoning methodology is not limited to confirming the results of the GeneCalling™ methodology, and can be equally applied to the confirming the results obtained from any other protocol utilizing nucleic acid species possessing the previously described generic structural motif. Therefore, as will be apparent to those of skill in the art, the oligo-poisoning confirmation methodology may be, more generally, utilized to confirm a putative sequence identification of a fragment within a sample of nucleic acid fragments possessing the aforementioned generic structural motif, that is, possessing known terminal subsequences of a length adequate to permit reliable and specific primer hybridization under stringent conditions for PCR amplification.

While several methods have been described in the art for the generation of such nucleic acid species from biological nucleic acid samples, it should be noted, however, that the Applicants do not hereby admit that any of the subsequently described examples contained herein are prior art to their invention. Three such exemplar methodologies will now be briefly described. A first method is disclosed in European Patent Application 0 534 858 Al, entitled "*Selective Restriction Fragment Amplification: A General Method for DNA Fingerprinting,*" and which is incorporated by reference herein in its entirety. According to this method, a sample of cDNA is initially digested with restriction endonucleases ("RE") into fragments and oligonucleotides complementary to these digested fragments are hybridized to the fragments. A longer primer strand of each adaptor is then ligated to the fragments. These products are then PCR amplified using PCR primers which include the longer primer strands. For selective amplification, these primers can, optionally, extend for 1-10 selected nucleotides beyond any remaining portion of the RE recognition site. Since fragments in the unamplified, amplified, and selectively amplified samples are all terminated by known primer sequences, this method generates nucleic acid samples of the described generic structure. In accord with this method, the

As previously discussed, oligo-poisoning confirmation is also adaptable to other methodologies which utilize nucleic acid fragment samples having the aforementioned generic structural motif which are either known within the art, or subsequently described in the future. As confirmatory oligo-poisoning methodologies are, preferably, applied to GeneCalling™ reaction products, they are described in the following subsection primarily with respect to such GeneCalling™ reaction products. However, this description is without limitation, as individuals possessing ordinary skill within the relevant arts will readily appreciate how to adapt oligo-poisoning methodologies to any sample of nucleic acids which possess the previously-described generic structural motif, including nucleic acid species produced by the aforementioned methods and the like.

(a) Conformation of a Putative Sequence by the Oligo-Poisoning Methodology

The oligo-poisoning methodology disclosed herein may be utilized to confirm a putative sequence which has been identified for a nucleic acid fragment, within a sample of nucleic acids, possessing the previously-described generic structural motif. The oligo-poisoning methodology depends upon the knowledge of, and serves to confirm the nucleotide sequence of, a portion of a unique, central nucleic acid sequence of interest, which is spatially located adjacent to known terminal subsequences. It has been ascertained that the knowledge of (at a minimum) the sequence of a portion of a fragment is, in fact, sufficient to confirm that a putative, candidate sequence, or which one of a small number of putative, candidate sequences, is actually the sequence of the nucleic acid species of interest.

For example, in the case of GeneCalling,™ fragments are, preferably, putatively identified by the computer-based GeneCalling™ analysis methods applied to the resultant GeneCalling™ signals. Even within complex genomes, it has been determined that the computer-based methods typically determine 1 or 2, usually less than 5, and almost invariably less than 10, potential "candidate" sequences for a particular GeneCalling™ signal. Accordingly, knowledge of only a few additional nucleotides of sequence is sufficient to verify which of the putative "candidate" sequences is actually the fragment producing the GeneCalling™ signal. Furthermore, such knowledge is also sufficient to differentiate known candidate sequences from previously uncharacterized nucleic acids.

aliquot is then separated, preferably, via gel electrophoresis, and the resultant separated bands are detected and analyzed in an appropriate manner (i.e., automated optical detection with the generation of an electropherogram). The results of the oligo-poisoning amplification reaction may then be compared with those results obtained from the original GeneCalling™ amplification reaction. This comparison allows any differences in the electrophoretic banding patterns and/or electrophoretic mobility of the nucleic acid fragments, especially any such differences in the band representing the fragment of interest, to be noted.

Accordingly, if the nucleic acid fragment of interest possesses a correctly identified putative sequence, *only* that band containing that nucleic acid fragment will either be absent or display altered electrophoretic mobility. The bands representing the other fragments will again be present to the same extent as found for the original amplification, electrophoretic separation, and detection performed without the "poisoning" primers. In contrast, if the fragment of interest possesses an incorrectly identified putative sequence, the band containing that fragment will, upon electrophoretic separation and detection, again be present to the same extent as found for the original amplification, electrophoretic separation, and detection performed without the "poisoning" primers. Therefore, in the incorrect identification scenario, the addition of the "poisoning" primers will have no demonstrable affect on the nucleic acid banding pattern obtained by electrophoretic separation of the PCR amplification reaction products.

In brief, the PCR amplification-mediated, oligo-poisoning methodology, as applied to GeneCalling™ confirmation, is comprised of the following steps:

Step 1: A PCR amplified GeneCalling™ reaction is performed.

Step 2: Utilizing the electrophoretic mobility results obtained from the electrophoresis of the GeneCalling™ PCR amplification reaction products in combination with those putative sequence "identity" results obtained from the utilization of the nucleic acid sequence database, a set of two "poisoning" oligonucleotide primers are designed, wherein each of the "poisoning" primers is complementary to one of the two RE initially utilized to digest the cDNA fragments (see Section 1; Step 2). The design of the "poisoning" primers incorporates: (*i*) sequence which is homologous to the RE-specific adapter sequence (See Section 1; Step 3)

fragments having expression differences between the samples (i.e., exhibiting differential expression), and to determine whether a novel nucleic acid is generating such expression differences.

For example, in the case of a fragment of interest which has been determined to be differentially expressed in each of two tissue samples (e.g., by a previous electrophoretic comparison) and which has been identified as possibly possessing two or more putative candidate sequences, the sequential "poisoning" of the fragments with two "poisoning" primers (each constructed to "poison" one of the two candidate sequences) may be utilized to identify the differential and relative presence of each candidate sequence within each tissue. In one potential scenario, the expression of both candidate sequences may be differentially increased within the same tissue sample, thus leading to a greater differential expression of the fragment of interest between the two tissues. In a second potential scenario, the expression of the candidate sequences may be differentially increased within different tissue samples, leading to a lesser differential of the fragment of interest. The oligo-poisoning methodology possesses the ability to ascertain which of these potential scenarios is correct.

(b) Preferred PCR Amplification Methodology Utilizing Oligo-Poisoning

The oligo-poisoning methodology is outlined below in detail first, with respect to construction of "poisoning" primers and second, with respect to the PCR amplification reaction conditions utilizing the "poisoning" primers. The following guidelines, described with respect to Figures. 3A and 3B, set forth the preferred criteria for the generation of "poisoning" primers. These figures illustrate the preferred application of oligo-poisoning to GeneCalling™ reactions and any differences which are appropriate for general application will be described as necessary. Figure 3A illustrates an exemplar dsDNA fragment ("fragment 1801")which is present in GeneCalling™ reaction products following adapter ligation and PCR amplification. Each strand of double-stranded fragment 1801 possesses: (*i*) a known 5'-terminal (and, therefore, a known complementary 3'-terminal) subsequences and (*ii*) a putatively identified central subsequence ("subsequence 1806") which is to be confirmed by utilization of the oligo-poisoning methodology.

The known 5'-terminal subsequence consists (on the upper DNA strand) of concatenated subsequences 1802 and 1803, and (on the lower DNA strand) of concatenated subsequences 1804 and 1805. Subsequences 1802 and 1804 have the same sequence as that of the adapter primers which were ligated onto the termini of the sample nucleic acid, thus generating fragment

adjacent portion of the 5'-terminal strand of central subsequence 1806, for which a partial or full candidate sequence has already been putatively identified. Hence, in view of recited structure of dsDNA fragment 1801, the "poisoning" primer is capable of annealing to the 3'-strand 1811 to facilitate PCR amplification of fragment 1801. The length of subsequence 1810 is chosen such that it reliably and specifically anneals to the complementary portion of strand 1811 under stringent hybridization conditions. The preferred, stringent hybridization conditions will be disclosed *supra*.

In the general case of nucleic acids species possessing the previously-described generic structural motif, subsequences 1809 is that of the 3'-terminus of the corresponding known terminal subsequence. Subsequence 1807 may be absent. Poisoning primer 1808 is preferably constructed according to the following specification in order for it to reliably and specifically recognize the putatively identified sequence for central subsequence 1806 under stringent hybridization conditions. For such stringent hybridization, subsequence 1810 is, preferably, 8-16 nt. in length, and is of sufficient length such that the 3'-terminus nucleotide 1812 of the "poisoning" primer is G or C. The most preferable length of subsequence 1810 is approximately 12 nucleotides. For reliable, specific and stringent hybridization, it is also preferable for the G+C content of subsequence 1810 to be at least approximately 40%, or more preferably from 50-60%, or greater. Therefore, a "poisoning" primer is preferably constructed and utilized for annealing to the terminus of the fragment containing the greatest overall G+C content. Additionally, the length of subsequence 1809 is such that the total length of "poisoning" primer 1808 is preferably between 18 and 30 nt. and most preferably between 19-23 nt. in order to facilitate reliable, specific, and stringent primer annealing. Where the length of remaining portion 1803 of the RE recognition site is comprised of 5 nt. and the length of 3'-terminus subsequence 1810 is 14 nt., 5'-terminus subsequence 1809, most preferably, possesses the same sequence as the 0-4 nt. long, 3'-terminus of adapter primer sequence 1802.

As previously discussed, a "poisoning" primer constructed pursuant to these aforementioned specifications is advantageously capable of specifically annealing to its complementary sequence under stringent hybridization conditions. For example, preferably, the melting temperature (T_m) of the "poisoning" primer is in the range from 5°C to 80°C, and more preferably, above 68°C. Such preferred T_m may be achieved, as is well-known within the art, by an appropriate G+C content or by the use of an appropriately long nucleotide sequence. For example, a preferable G+C content is from 40-60%. Therefore, where the composition of the 5'-terminus of central subsequence 1806 includes a higher percentage of A+T, subsequence 1810

the concentration of nucleic acids in the sample must be such that PCR amplification does not saturate and allow residual fragments from the input sample to obscure subsequent separation and detection. Alternately, the biotin clean-up procedure described can be used to eliminate residual fragments from the input sample. Where nucleic acid fragment samples have been previous amplified, as in the GeneCalling™ method, the samples are preferably diluted before PCR amplification with the "poisoning" primer in order that amplified fragments can be clearly distinguished from any residual fragments left from the input sample. Such a dilution is preferably at least 1:50 (v/v), and is more preferably 1:100 (v/v) in order to reduce residual fragment concentration to an approximately 1% or less background level, and can be 1:1000 (v/v) or greater in order to resolve especially ambiguous or low concentration fragments. In addition, in order that the amplification of all fragments with a candidate identified sequence is substantially solely due to the "poisoning" primers and not due to the adapter primers, the "poisoning" primers are preferably present in a molar excess to the regular adapter primers. This molar excess is preferably at least 1:50, and is more preferably 1:100 in order to reduce amplification of fragment having the "poisoned" sequence to an approximately 1% or less background level, and can be 1:1000 or greater in order to resolve especially ambiguous or low concentration fragments.

Generally, other parameters of the PCR reactions are preferably similar or identical to those used in the generation of GeneCalling™ signals. This is especially advantageous in the case of application of oligo-poisoning to GeneCalling™ because poisoned signals can be readily compared to the initial GeneCalling™ signals. Such PCR parameters are advantageous also in the case of oligo-poisoning applied to nucleic acid samples produced according to other methods. Oligo-poisoning also is adaptable to other high-stringency PCR protocols known in the art. Details of the preferred, exemplar PCR protocol will be disclosed in a subsequent section. In particular, it is preferred that a "hot-start" PCR method be used, and this preferred "hot-start" method also include the use of the wax layering technique described *supra*.

In this application of this wax technique, PCR reaction vessels are set up by placing dNTPs and water in the lower portion of a reaction vessel; layering wax on top of this dNTP solution; and placing the remainder of the PCR reaction mix on top on the wax layer. As previously described, the wax used preferably melts rapidly at near but less than 72°C, the temperature preferred for the extension phase of the PCR amplification. During PCR amplification, the first thermal cycle begins with a denaturing temperature of approximately 96°C, which is adequate to melt the wax, cause mixing of the reagent compartments, and initiate

the PCR reaction in place of dGTP. This nucleotide analog has been shown to increase PCR efficiency for G+C rich targets. See *e.g.*, Mutter, *et al.*, 1995. *Nuc. Acid Res.* 23:1411-1418). Another such technique is the addition of tetramethylammonium chloride to the reaction mixture, which has the effect of raising the T_m. See *e.g.*, Chevet, *et al.*, 1995, *Nuc. Acids Res.* 23:3343-3344.

The PCR temperature profile is performed according to the preferred protocol for a certain number of cycles. Following the amplification step, optional cleanup and separation steps prior to length separation and fragment detection can be advantageous to substantially eliminate certain unwanted DNA strands and thereby to improve the signal to noise ratio of GeneCalling™ signals, or to substantially separate the reaction products into various classes and thereby to simplify interpretation of detected fragment patterns by removing signal ambiguities. For example, unused primer strands and single strands produced by linear amplification are unwanted in later steps. These steps are based upon various types of primer enhancements including conjugated capture moieties and release means.

In one embodiment of these optional primer enhancement steps where one of the two primers used has a conjugated capture moiety, GeneCalling™ reaction products fall into certain categories. These categories (described without limitation in the case where the capture moiety is biotin) include:

(a) dsDNA fragments neither strand of which has a biotin moiety;
(b) dsDNA fragments having only one strand with a conjugated biotin moiety;
(c) dsDNA molecule fragments having biotin moieties conjugated to both strands; and
(d) unwanted single-stranded DNA (ssDNA) strands with and without conjugated biotin.

The additional method steps comprise contacting the amplified fragments with streptavidin affixed to a solid support, preferably streptavidin magnetic beads, washing the beads to in a non—denaturing wash buffer to remove unbound DNA, and then resuspending the beads in a denaturing loading buffer and separating the beads from this buffer. The denatured single strands are then passed to the separation and detection steps.

As a results of these steps only the strand of category "b" without biotin is removed in the loading buffer for separation and detection. Thereby, only fragments cut on either end by different REs and freed from single stranded contaminants are separated and detected with minimized noise. Category "a" products are not bound to the beads and are washed away in the

Biosystems, Inc.) for analysis. The electrophoresis can be done by suspending the reaction products in a loading buffer, which can be non-denaturing, in which the dsDNA remains hybridized and carries the labels (if any) of both primers. The buffer can also be denaturing, in which the dsDNA separates into single strands that typically are expected to migrate together (in the absence of large average differences in strand composition or significant strand secondary structure).

The length distribution is detected with various detection means. If no labels are used, means such as antigen (Ag) and antibody (Ab) staining and intercalating dyes can be used. Here, it can be advantageous to separate reaction products into classes, according to the previously described protocols, in order that each band can be unambiguously identified as to its target end subsequences. In the case of fluorochrome labels, since multiple fluorochrome labels can be typically be resolved from a single band in a gel, the products of one recognition reaction with several REs or other recognition means or of several separate recognition reactions can be analyzed in a single lane. However, where one band reveals signals from multiple fluorochrome labels, interpretation can be ambiguous: is such a band due to one fragment cut with multiple REs or to multiple fragments each cut by one RE. In this case, it can also be advantageous to separate reaction products into classes.

Following detection, the resulting electrophoretic banding patterns and mobilities are utilized to generate an electropherogram. The electropherogram provides a graphical plot of the electrophoretic mobilities of each individual, amplified nucleic acid fragment.

3. Utilization of the Oligo-Poisoning Methodology for Sequence Confirmation of the Human C1r Gene Sequence

The application of the oligo-poisoning methodology to the analysis of a RE-generated sequence derived from the Human Complement Component 1, Subcomponent r (c1R) gene will now be discussed.

Human c1R poly(A)$^+$ mRNA was utilized to generate a homologous cDNA by standard protocols known within the relevant fields. Figure 4 illustrates the nucleotide sequence of the 2493 c1R cDNA. Following synthesis, the c1R cDNA was digested with the REs BspH1 (recognition sequence - TCATA) and EcoR1 (recognition sequence - GAATTC) to produce a 318 bp. fragment. With reference to Figure 4, the recognition sites for BspH1 and EcoR1 are shown by the underlined sequence; whereas the nucleotide sequence of the 318 bp. fragment

amplification, electropherograms were generated for the amplification reactions utilizing the two "poisoning" primers. Figure 6, Panels A & B illustrate the electropherograms for the "poisoning" primer corresponding to the J-primer and the 'poisoning" primer corresponding to the R-primer, respectively. Comparison of Figure 6, Panels A & B, demonstrates an approximate 3-fold reduction in the signal corresponding to the 318.7 nt. fragment. Thus, confirming the nucleic acid sequence identity derived from the original GeneCalling™ methodology.

The present invention is not to be limited in scope by the specific embodiments disclosed herein. Indeed, various modifications of the present invention, in addition to those described herein, will become readily apparent to those individuals skilled in the relevant arts from the foregoing descriptions and accompanying figures. Such modifications are intended to fall within the scope of the appended claims. In addition, various publications are cited herein and their disclosures are hereby incorporated by reference in their entirety.

3. The method of claim 1 wherein one of said generated signals does not have a sequence in said sequence database determined to match said signal, and the confirming step (d) additionally comprises the steps of:

(a) recovering a fragment of a nucleic acid in the sample which generates said signal;

(b) sequencing said fragment to determine at least a partial sequence for said fragment; and

(c) ~~confirming~~ verifying that said sample comprises a nucleic acid having a sequence comprising at least a portion of said determined sequence.

4. The method of claim 1 wherein said plurality of nucleic acids are DNA and wherein the step of probing comprises:

(a) digesting the sample with one or more restriction endonucleases, said restriction endonucleases having recognition sites that are said target subsequences and leaving single-stranded nucleotide overhangs on the digested ends;

(b) hybridizing double-stranded adapter nucleic acids with the digested sample fragments, said adapter nucleic acids having an end complementary to one of said single—stranded overhangs; and

(c) ligating with a ligase a strand of said adapter nucleic acids to the 5'-end of a strand of the digested sample fragments to form ligated nucleic acid fragments.

8. The method of claim 7 wherein said regular primers are labeled and said poisoning primer labeled such that it is distinguishable from said regular primers upon detection.

9. The method of claim 7 wherein said regular primers are labeled and said poisoning primer is unlabeled.

10. The method of claim 9 further comprising before step (a) the steps of:

(i) contacting said nucleic acid fragments in said sample in amplifying conditions with a nucleic acid polymerase and said regular primer oligonucleotides, said regular primer oligonucleotides being detectably labeled;

(ii) separating the products of step (ii);

(iii) detecting said separated products; and wherein step (c) further comprises confirming said putatively identified sequence if said nucleic acid fragment with a putatively identified sequence is detected in step (iii) and not detected in step (c).

11. The method of claim 9 wherein step (c) further comprises confirming said putatively identified sequence if said nucleic acid fragment with a putatively identified sequence is not detected.

12. The method of claim 7 wherein said sample of nucleic acids is produced according to said probing and said generating steps of claim 1.

13. The method of claim 7 wherein said sample of nucleic acids is produced according to a method comprising:

(a) synthesizing cDNA from mRNA using a method comprising contacting said mRNA with a first-strand primer comprising one or more phasing nucleotides and a first non-complementary heel subsequence;

(b) digesting said cDNA into fragments with a restriction endonuclease;

(c) ligating to said fragments a partially double stranded second primer comprising a second 5' non-complementary subsequence; and

(d) amplifying said ligated fragments by contacting them with a DNA polymerase and a first amplifying primer comprising at least a portion of said first non-

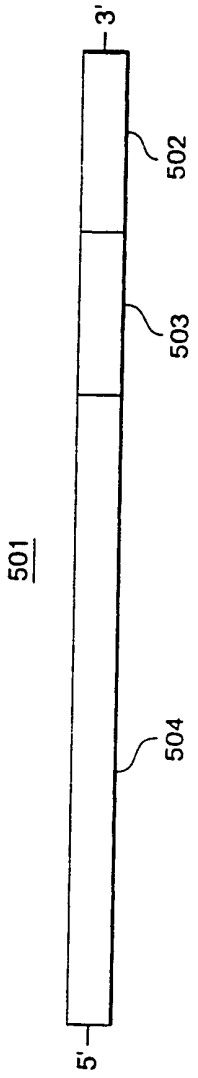

Fig. 1

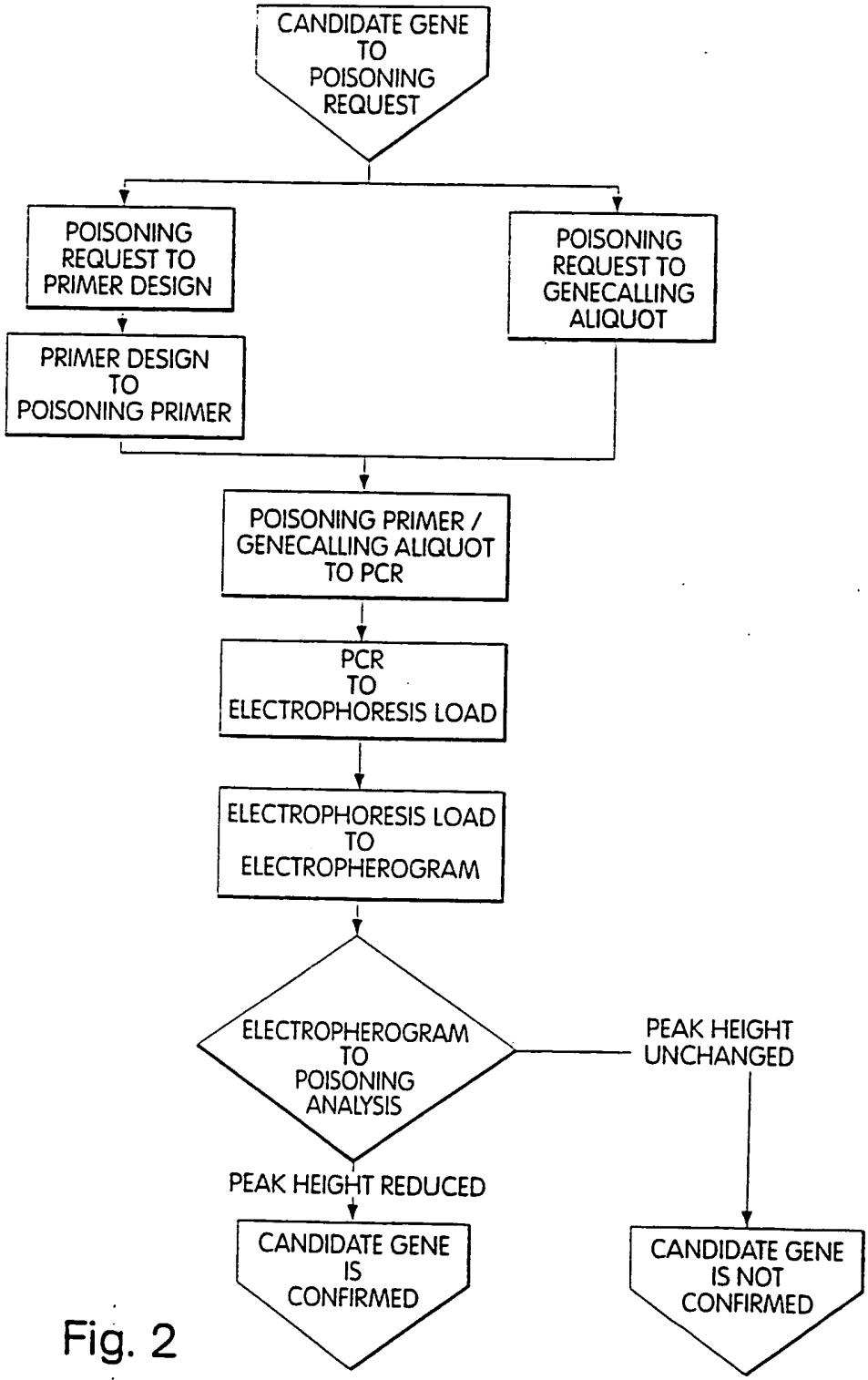
CANDIDATE GENE TO POISONING REQUEST
POISONING REQUEST TO PRIMER DESIGN
POISONING REQUEST TO GENECALLING ALIQUOT
PRIMER DESIGN TO POISONING PRIMER
POISONING PRIMER / GENECALLING ALIQUOT TO PCR
PCR TO ELECTROPHORESIS LOAD
ELECTROPHORESIS LOAD TO ELECTROPHEROGRAM
ELECTROPHEROGRAM TO POISONING ANALYSIS
PEAK HEIGHT UNCHANGED
PEAK HEIGHT REDUCED
CANDIDATE GENE IS CONFIRMED
CANDIDATE GENE IS NOT CONFIRMED

Fig. 2

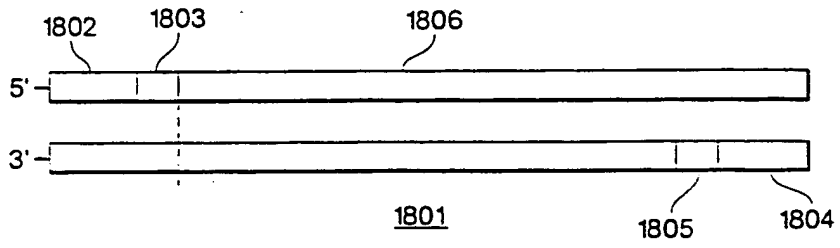
1802
1803
1806
5'
3'
1801
1805
1804

Fig. 3A

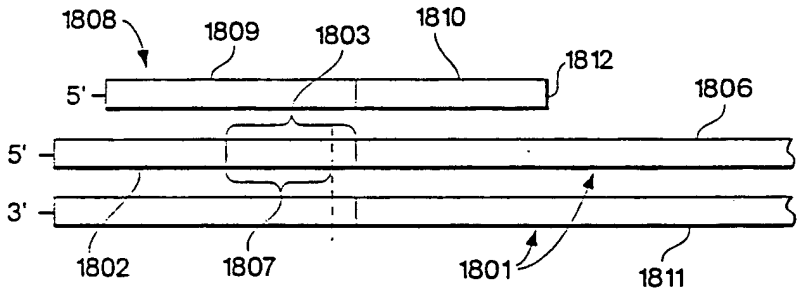
1808
1809
1803
1810
1812
5'
1806
5'
3'
1802
1807
1801
1811

Fig. 3B

- HOMO SAPIENS, 2493 bp (RNA).

GENE SEQUENCE

```
   1 ggatcgattt gagtaagagc atagctgtcg ggagagccca ggattcaaca cgggccttga
  61 gaaatgtggc tcttgtacct cctggtgccg gccctgttct gcagggcagg aggctccatt
 121 cccatccctc agaagttatt tggggaggtg acttcccctc tgttccccaa gccttacccc
 181 aacaactttg aaacaaccac tgtgatcaca gtccccacgg gatacagggt gaagctcgtc
 241 ttccagcagt ttgacctgga gccttctgaa ggctgcttct atgattatgt caagatctct
 301 gctgataaga aaagcctggg gaggttctgt gggcaactgg gttctccact gggcaacccc
 361 ccgggaaaga aggaatttat gtcccaaggg aacaagatgc tgctgacctt ccacacagac
 421 ttctccaacg aggagaatgg gaccatcatg ttctacaagg gcttcctggc ctactaccaa
 481 gctgtggacc ttgatgaatg tgcttcccgg agcaaatcag gggaggagga tccccagccc
 541 cagtgccagc acctgtgtca caactacgtt ggaggctact tctgttcctg ccgtccaggc
 601 tatgagcttc aggaagacag gcattcctgc caggctgagt gcagcagcga gctgtacacg
 661 gaggcatcag gctacatctc cagcctggag taccctcggt cctacccccc tgacctgcgc
 721 tgcaactaca gcatccgggt ggagcggggc ctcaccctgc acctcaagtt cctggagcct
 781 tttgatattg atgaccacca gcaagtacac tgcccctatg accagctaca gatctatgcc
 841 aacgggaaga acattggcga gttctgtggg aagcaaaggc cccccgacct cgacaccagc
 901 agcaatgctg tggatctgct gttcttcaca gatgagtcgg gggacagccg gggctggaag
 961 ctgcgctaca ccaccgagat catcaagtgc ccccagccca agaccctaga cgagttcacc
1021 atcatccaga acctgcagcc tcagtaccag ttccgtgact acttcattgc tacctgcaag
1081 caaggctacc agctcataga ggggaaccag gtgctgcatt ccttcacagc tgtctgccag
1141 gatgatggca cgtggcatcg tgccatgccc agatgcaaga tcaaggactg tgggcagccc
1201 cgaaacctgc ctaatggtga cttccgttac accaccacaa tgggagtgaa cacctacaag
1261 gcccgtatcc agtactactg ccatgagcca tattacaaga tgcagaccag agctggcagc
1321 agggagtctg agcaaggggt gtacacctgc acagcacagg gcatttggaa gaatgaacag
1381 aagggagaga agattcctcg gtgcttgcca gtgtgtggga agcccgtgaa ccccgtggaa
1441 cagaggcagc gcataatcgg agggcaaaaa gccaagatgg gcaacttccc ctggcaggtg
1501 ttcaccaaca tccacgggcg cgggggcggg gccctgctgg gcgaccgctg gatcctcaca
1561 gctgcccaca ccctgtatcc caaggaacac gaagcgcaaa gcaacgcctc tttggatgtg
1621 ttcctgggcc acacaaatgt ggaagagctc atgaagctag gaaatcaccc catccgcagg
1681 gtcagcgtcc acccggacta ccgtcaggat gagtcctaca attttgaggg ggacatcgcc
1741 ctgctggagc tggaaaatag tgtcaccctg ggtcccaacc tcctccccat ctgcctccct
1801 gacaacgata ccttctacga cctgggcttg atgggctatg tcagtggctt cggggtcatg
1861 gaggagaaga ttgctcatga cctcaggttt gtccgtctgc ccgtagctaa tccacaggcc
1921 tgtgagaact ggctccgggg aaagaatagg atggatgtgt tctctcaaaa catgttctgt
1981 gctggacacc catctctaaa gcaggacgcc tgccaggggg atagtggggg cgtttttgca
2041 gtaagggacc cgaacactga tcgctgggtg gccacgggca tcgtgtcctg gggcatcggg
2101 tgcagcaggg gctatggctt ctacaccaaa gtgctcaact acgtggactg gatcaagaaa
2161 gagatggagg aggaggactg agcccagaat tcactaggtt cgaatccaga gagcagtgtg
2221 gaaaaaaaaa aaacaaaaaa caactgacca gttgttgata accactaaga gtctctatta
2281 aaattactga tgcagaaaga ccgtgtgtga aattctcttt cctgtagtcc cattgatgta
2341 ctttacctga aacaacccaa aggccccttt ctttcttctg aggattgcag aggatatagt
2401 tatcaatctc tagttgtcac tttcctcttc cactttgata ccattgggtc attgaatata
2461 acttttttcca aataaagttt tatgagaaat gc
```

Fig. 4

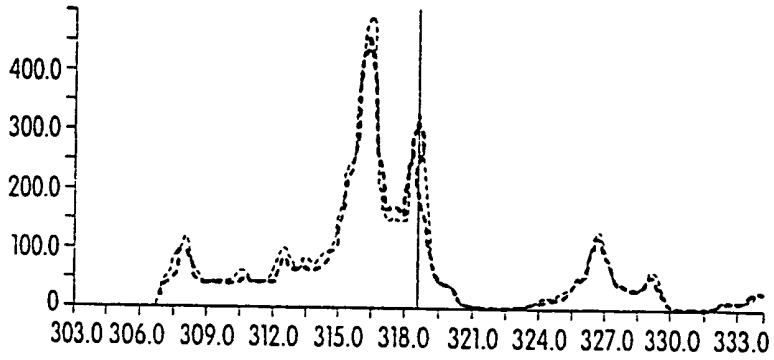
400.0
300.0
200.0
100.0
0
303.0 306.0 309.0 312.0 315.0 318.0 321.0 324.0 327.0 330.0 333.0

Fig. 5A

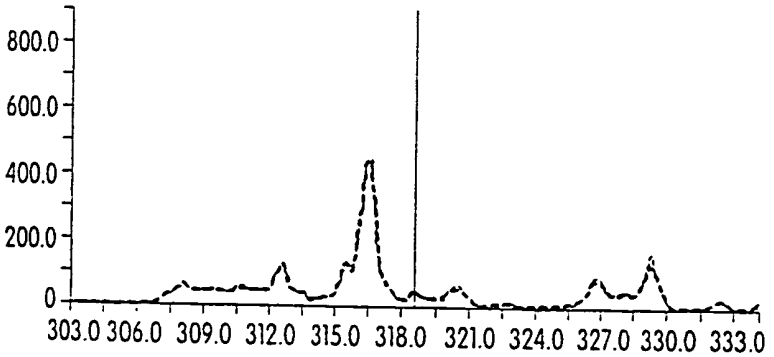
800.0
600.0
400.0
200.0
0
303.0 306.0 309.0 312.0 315.0 318.0 321.0 324.0 327.0 330.0 333.0

Fig. 5B

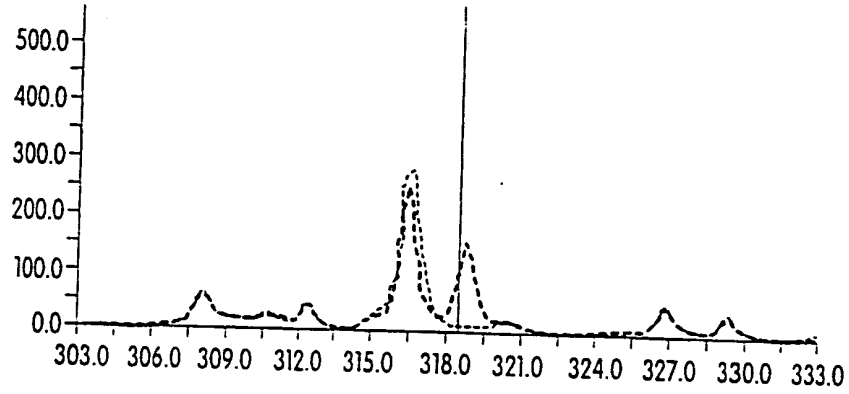
500.0
400.0
300.0
200.0
100.0
0.0
303.0 306.0 309.0 312.0 315.0 318.0 321.0 324.0 327.0 330.0 333.0

Fig. 6A

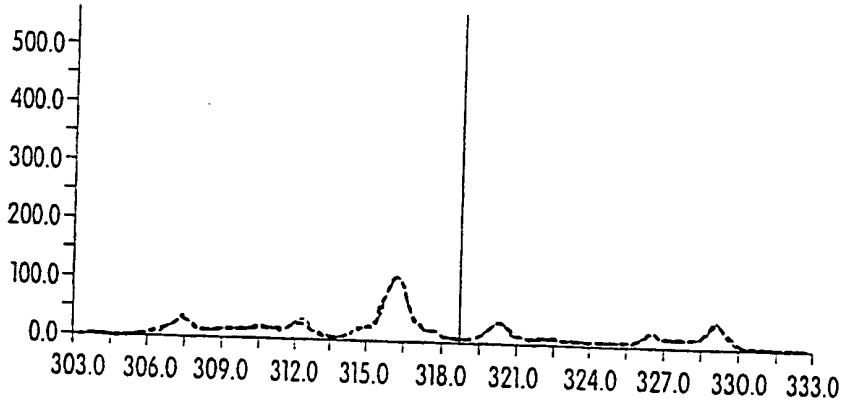
500.0
400.0
300.0
200.0
100.0
0.0
303.0 306.0 309.0 312.0 315.0 318.0 321.0 324.0 327.0 330.0 333.0

Fig. 6B